MEDICAL UNIVERSITY – PLEVEN
FACULTY OF PUBLIC HEALTH

DEPARTMENT OF PUBLIC HEALTH SCIENCES
CENTRE FOR DISTANT LEARNING

LECTURE No5

CORRELATION. CORRELATION COEFFICIENTS FOR QUANTITATIVE AND QUALITATIVE DATA

Assoc. Prof. Gena Grancharova, MD, PhD

Plan of the lecture

1. Definition of basic concepts
2. Types of correlation
3. Correlation coefficients
4. Uses of correlation in health sciences
5. Correlation and causation

1. BASIC CONCEPTS

- A fundamental aim of scientific and clinical research is to establish relationships between two or more sets of observations or variables.

- Finding such relationships is often a fundamental initial step for identifying causal relationships.

- **The correlation is concerned with expressing quantitatively the degree and direction of the relationship between variables.**

- **Correlation is useful in the health sciences in areas such as determining the validity and reliability of clinical measures or in expressing how health problems are related to crucial biological, behavioural or environmental factors.**

Consider the following statements:

- **1. There is a positive relationship between cigarette smoking and lung damage.**

- This statement is implying that there is evidence that if you score high on one variable (cigarette smoking) you are likely to score high on the other variable (lung damage).

- **2. There is a negative relationship between being overweight and life expectancy.**

- The second statement describes the finding that scoring high on the variable 'overweight' tends to be associated with low scores on the variable 'life expectancy'.

- The information missing from each of the statements is the numerical value for the degree or magnitude of the association between variables, e.g. **how strong is the association.**

2. Types of correlation

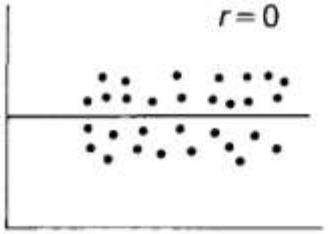
In general, a variety of relationships are possible between two variables:

- **1. Linear correlation:**

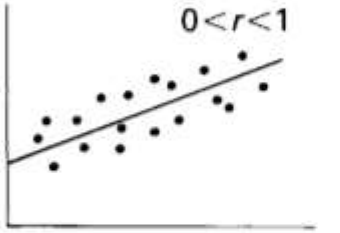
Positive - high scores on x are related to high scores on y;

Negative - high scores on x are associated with low scores on y;

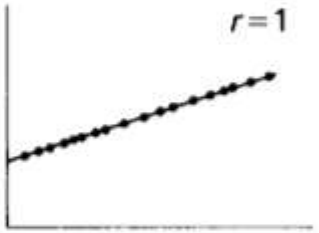
- **2. Non-linear correlation** - the relationship between x and y is represented by a curve



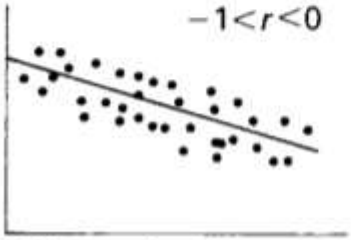
(a) No correlation



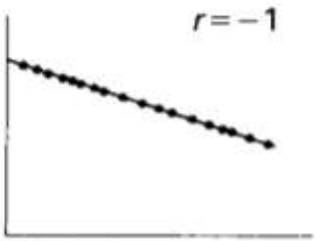
(b) Imperfect positive correlation



(c) Perfect positive correlation

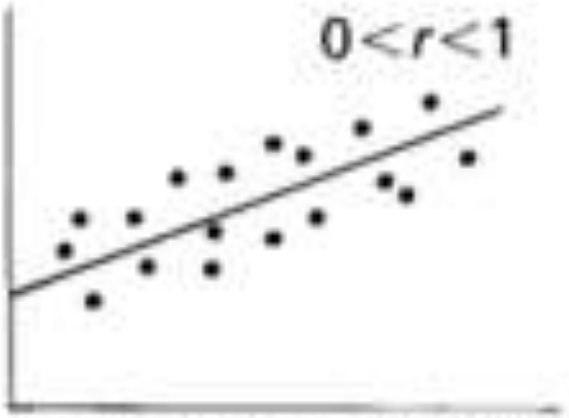


(d) Imperfect negative correlation

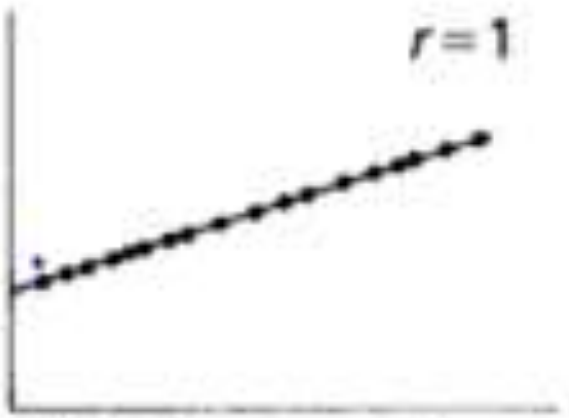


(e) Perfect negative correlation

Scatter plots illustrating different types of correlation

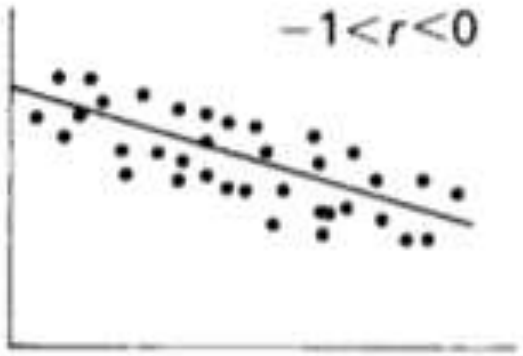


(b) Imperfect positive correlation

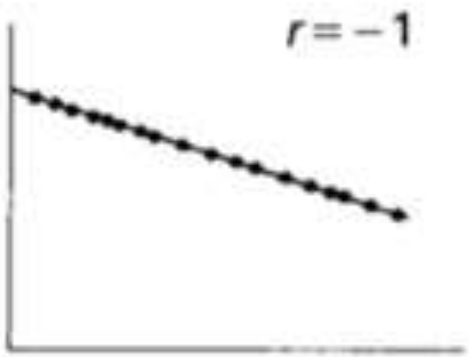


(c) Perfect positive correlation

The graph represents a **positive correlation**, indicating that high scores on x are related to high scores on y. For instance, the relationship between cigarette smoking and lung damage is a positive correlation.



(d) Imperfect negative correlation



(e) Perfect negative correlation

The graph represents a **negative correlation**, where high scores on x are associated with low scores on y. For instance, the correlation between the variables 'being overweight' and 'life expectancy' is negative, meaning that the more you are overweight, the lower your life expectancy.

There is no correlation at all.

The graph represents a **non-linear correlation,** where a curve best represents the relationship between x and y.

3. CORRELATION COEFFICIENT

- **A correlation coefficient** is a descriptive statistic that expresses the degree or the magnitude of the association between the variables and the direction of the correlation.

- In order to demonstrate that two variables are correlated, we must obtain measures on both variables for the same set of subjects or events.

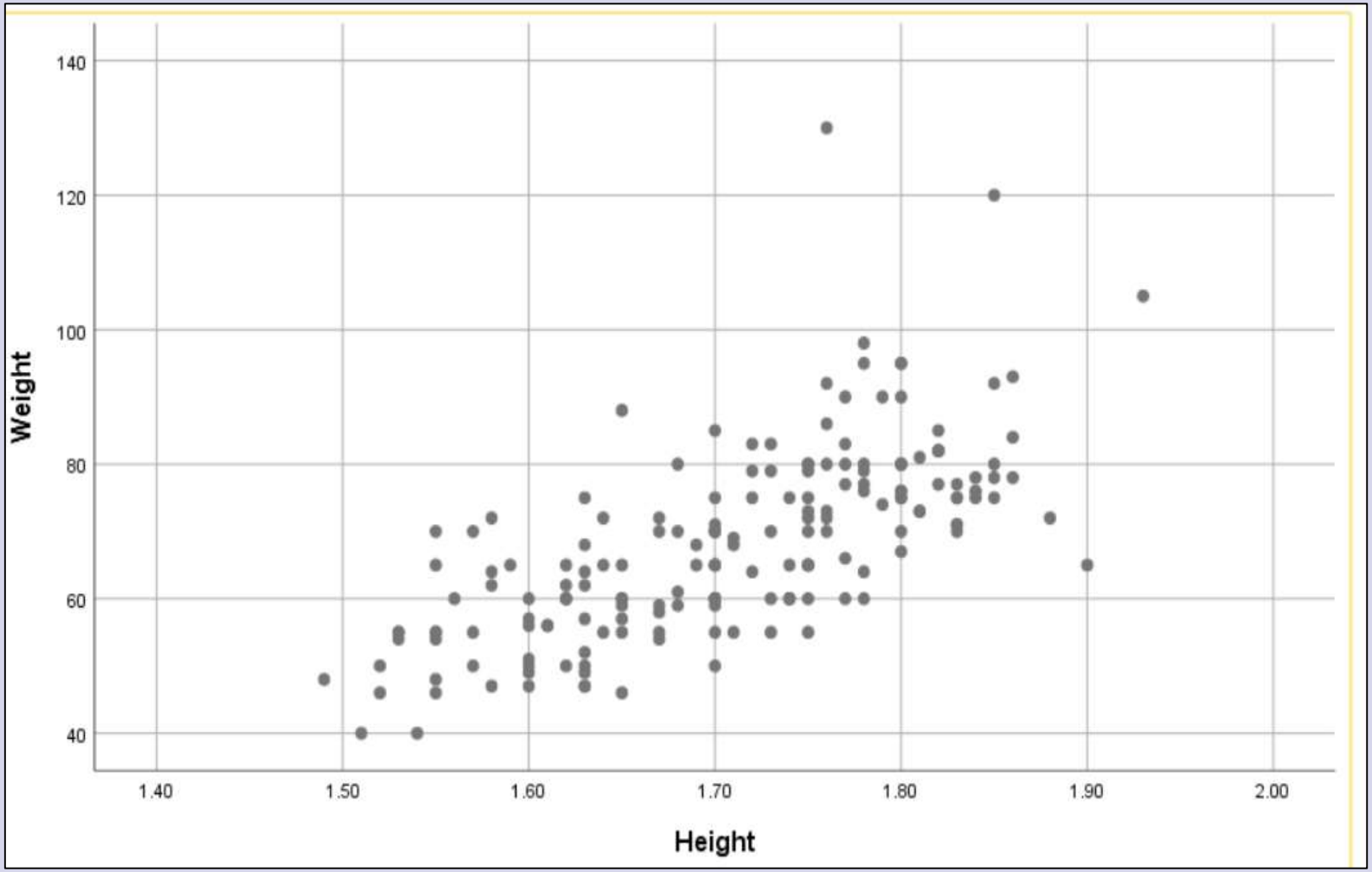
CORRELATION COEFFICIENT

- For a visual representation of the relationship between two variables, we can plot the data on a **scattergram.**

- **A scattergram** is a graph of the paired scores for each subject on two variables, called by convention **x** and **y**.

CORRELATION COEFFICIENT

- For example, we can plot **a scattergram** with the variables “height” and “weight” based on the results in the database that we created for 185 students. We will find out that there is a **positive relationship** between the two variables. That means, students who have high scores for “height” (variable x) tend to have high scores for “weight” (variable y).



The scattergram represents a positive correlation between height and weight for 185 students attending the course in Medical statistics

CORRELATION COEFFICIENTS

- **When we need to know or express the numerical value of the correlation between variables x and y, we calculate a statistic called a correlation coefficient.**
- **The correlation coefficient expresses quantitatively the magnitude and direction of the correlation.**

CORRELATION COEFFICIENTS

Characteristics of correlation coefficients:

- 1. Correlation coefficients are calculated from pairs of measurements on variables x and y for the same group of individuals.

- 2. A positive correlation is denoted by (+) (plus sign) and a negative correlation - by (-) (minus sign).

CORRELATION COEFFICIENTS

3. The values of the correlation coefficients range from +1 to -1:

- **+1 means a perfect positive correlation;**
- **0 means no correlation at all;**
- **-1 means a perfect negative correlation.**

4. The square of the correlation coefficient is called a coefficient of determination.

Degrees of correlation

- **3-point scale**
  - **Over 0.7 - high**
  - **0.3 to 0.7 - moderate**
  - **less than 0.3 - weak**
- **5-point scale**
  - **0.00-0.25 - little, if any**
  - **0.26-0.49 - low**
  - **0.50-0.69 - moderate**
  - **0.70-0.89 - high**
  - **0.90-1.00 - very high**

CORRELATION COEFFICIENTS

SELECTION OF CORRELATION COEFFICIENT

- **There are several types of correlation coefficients used in statistics under specific conditions.**

CORRELATION COEFFICIENTS

Coefficient	Conditions where they are appropriate
φ (phi)	Both x and y are measured on a nominal scale
Sperman's ρ (rho)	Both x and y are measured on, or transformed to, ordinal scales
Pierson's r	Both x and y are measured on an interval or ratio scale

CORRELATION COEFFICIENTS

- **All the correlation coefficients shown in table above are appropriate for quantifying linear relationships between variables.**

- **There are other correlation coefficients, such as η (eta) which are used for quantifying non-linear relationships.**

Statistical methods for calculation the correlation coefficients can be classified according to the variables studied into four main groups:

1. The variables are qualitative and each of them has only two categories, e.g. in 2 x 2 tables – φ (phi);

2. The variables are qualitative with more than two categories, e.g. in multiple tables - φ (phi)

3. When both x and y are measured on, or transformed to, ordinal scales – Sperman correlation coefficient;

4. When both x and y are quantitative and measured on interval or ratio scale – rank correlation - Pierson coefficient

Pearson's r

- 1. When x and y are measured on an interval or a ratio scale;
- 2. Both variables are normally distributed;
- 3. It describes a linear relationship.

- Pearson's r is a measure of the extent to which paired scores are correlated. It is convenient for small pairs of scores.

Correlations between height and weight

		Weight	Height
Weight	Pearson Correlation	1	,686**
	Sig. (2-tailed)		,000
	N	185	185
Height	Pearson Correlation	,686**	1
	Sig. (2-tailed)	,000	
	N	185	185

**. Correlation is significant at the 0.01 level (2-tailed).

Coefficient of determination $r^2 = 0.47$ or 47%

! This example underlines that when both variables x and y are interchanged Pearson correlation r takes the same value.

Correlations between marks in Anatomy and Physiology

		Physiology	Anatomy
Physiology	Pearson Correlation	1	,499**
	Sig. (2-tailed)		,000
	N	185	185
Anatomy,	Pearson Correlation	,499**	1
	Sig. (2-tailed)	,000	
	N	185	185

**. Correlation is significant at the 0.01 level (2-tailed).

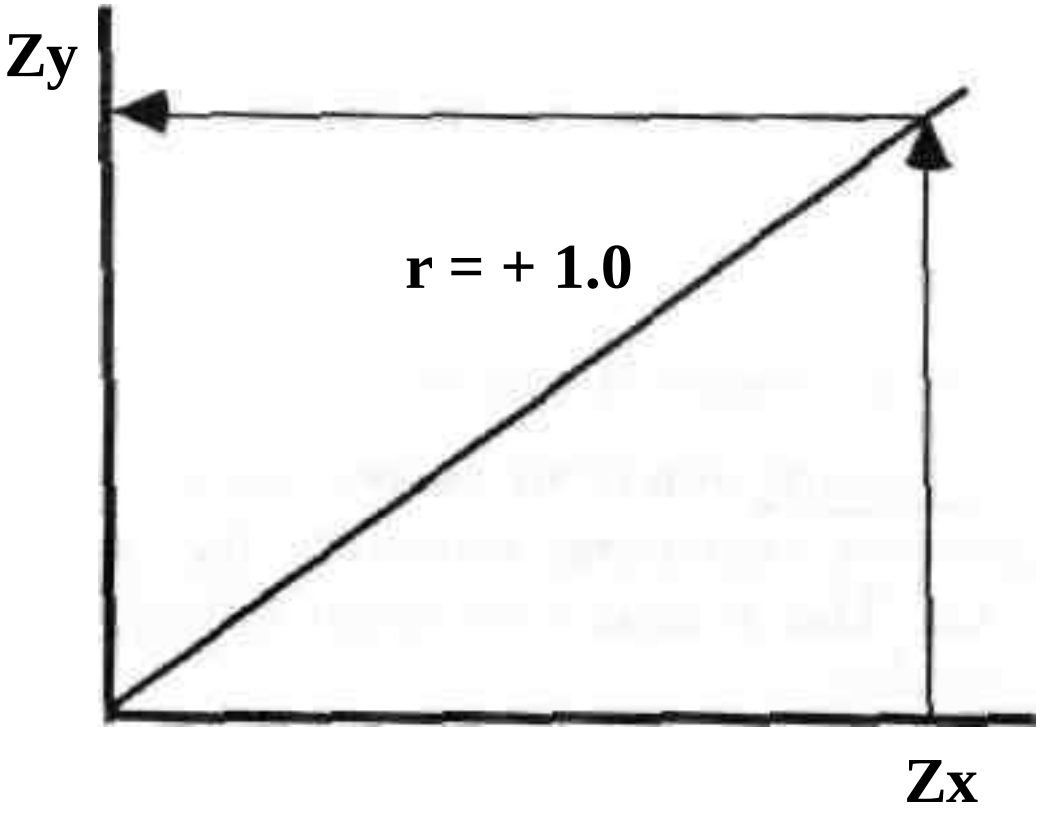
Coefficient of determination $r^2 = 0.249$ or 24,9%

! This example underlines that when both variables x and y are interchanged Pearson correlation r takes the same value.

4. Uses of correlation in health sciences

- **1. For Prediction** - when the correlation is very high or perfect. Given any score on x, we can transform it into z_x, then using the graph we can read off the corresponding z_y, and transforming z_y it into y we predict the value of y given the value of x.

Uses of correlation in health sciences



We can see that given any score on y we can transform this into a z score (zx) and then using the graph we can read off the corresponding score on y *(Zy)*. Of course it is extremely rare that there should be a perfect *(r* = +1) correlation between two variables. The smaller the correlation coefficient, the greater the probability in making an error in prediction.

Uses of correlation in the health sciences

- **2. For measuring the reliability and predictive validity of assessment**
- **Reliability - the extent to which a test or measurement result is reproducible. Reliability refers to measurements using instruments or subjective judgements remaining relatively the same on repeated administration. This is called test-retest reliability. The correlation coefficient can be used also to determine the degree of interobserver reliability.**
- **The higher the correlation, the greater the reliability. If the measurements are on interval or ratio data, we would have calculated Pearson's *r* to represent quantitatively the reliability of the measurement.**

Uses of correlation in the health sciences

- **Predictive validity** - **the extent to which a test or measure can validly predict a future event. For instance, say that we devise an assessment procedure to predict how much people will benefit from a rehabilitation programme. If the correlation between the assessment and a measure of the success of the rehabilitation programme is high (say 0.7 or 0.8), then the assessment procedure has high predictive validity. If, however, the correlation is low (say 0.4 or 0.3), then the predictive validity of the assessment is low, and it would be unwise to use the assessment as a screening procedure for the entry of patients into the rehabilitation programme.**

Uses of correlation in the health sciences

- **3. For estimating shared variance - we can use the square of r which is called a coefficient of determination - r^2, and represents the proportion of variance in one variable accounted for by the other.**
- **From the above example we have found out Pierson's r between x (height) and y (weight) equal to 0.686.**
- **r^2 = 0.686 x 0,686 = 0.47 or 47%**
- **This means that 47% of the variability of weight can be accounted for by height. The other 51% would be accounted for by other factors.**

Uses of correlation in the health sciences

- **We can also calculate the coefficient of determination between the marks in Anatomy (variable x) and marks in Physiology (variable y).**
- **Pierson's r is equal to 0.499.**
- **r^2 = 0.499 x 0.499 = 0.249 or about 25%**
- **This means that only 25% of the variability of marks in Anatomy can be accounted for by marks in Physiology. The other 75% would be accounted for by other factors.**

5. CORRELATION AND CAUSATION

- We must be very cautious in causal interpretation of correlation coefficients.
- A strong association between variables is an important step towards establishing causal links; and, although not sufficient without control it can help to discount plausible alternative hypotheses.
- Correlation is not the only criteria for establishing causal relationship.

CORRELATION AND CAUSATION

- There is often multiple causation in some health problems. So, we need to distinguish between competing plausible hypotheses.
- We need to establish the statistical significance of the correlation coefficient r.
- As with other descriptive statistics, caution is necessary when correlation coefficients are calculated for a sample and then generalized to a population. We must be certain that there is no selection or other type of bias.

TEST EXAMPLES

1. Correlation is defined as the relative difference between two variables.

A. True B. False

2. The association between two variables can be plotted on a scattergram.

A. True B. False

3. If the distribution of paired scores is best represented by a curve, the relationship is non-linear.

A. True B. False

4. When we speak of a positive (+) relationship, we mean that high scores on one variable are associated with high scores on the other variable.

A. True B. False

5. In a negative (-) relationship, low scores on one variable are associated with low scores on the other.

A. True B. False

6. There are several types of correlation coefficients, the selection of which is determined by the level of scaling of the two variables.

A. True B. False

7. When both variables are measured on an interval or ratio scale, Pearson's r is the most appropriate correlation coefficient.

A. True B. False

8. When both variables are measured on, or converted to, ordinal scales, we must use ϕ (phi) to express correlation.

A. True B. False

9. For two variables measured on nominal scales, we use ρ (rho) to express correlation.

A. True B. False

10. When we use Pearson's r, we assume that both variables are continuous and normally distributed.

A. True B. False

11. The calculated values of correlation coefficients range between 0 and –1.

A. True B. False

12. A correlation coefficient of –1.0 represents a very low linear correlation.

A. True B. False

13. The coefficient of determination is the square of the correlation coefficient.

A. True B. False

14. If r = 0.3, then the coefficient of determination will be 9.0.

A. True B. False

15. Say $r^2 = 0.36$ for a set of data. This implies that 36% of the variability of y is explained in terms of x.

A. True B. False

16. Say $r^2 = 0.48$ for a set of data. This implies that 52% of the variability of y is not explained in terms of x.

A. True B. False

17. Even a high correlation is not necessarily indicative of a causal relationship between two variables.

A. True B. False

18. As the value of r increases, the proportion of variability of y that can be accounted for by x, decreases.

A. True B. False

19. A scattergram is used to help to decide if the relationship between two variables is linear or curvilinear.

A. True B. False

20. Spearman's ρ (rho) is used when one or both variables are at least of interval scaling.

A. True B. False

21. A scattergram:
A. is a statistical test
B. must be linear
C. must be curvilinear
D. is a graph of x and y scores

22. If the relationship between x and y is positive, then as variable x decreases, variable y:
A. increases
B. decreases
C. remains the same
D. changes linearly

23. In a 'negative' relationship:
A. as x increases, y increases
B. as x decreases, y decreases
C. as x increases, y decreases

24. Which of the following correlation coefficients reflects the lowest strength of association?

A. –0.60

B. –0.33

C. 0.29

25. Which of the following correlation coefficients reflects the highest strength of association?

A. – 1.0

B. – 0.95

C. 0.85

26. Which of the following statements is false?

A. Spearman's ρ (rho) is used when one or both variables are at least of interval scaling.

B. The range of correlation coefficient is from –1 to +1.

C. A correlation of r = 0.85 implies a stronger association than r = -0.70

27. You are told there is a high inverse association between the variables ‘amount of exercise’ and ‘incidence of heart disease’. The correlation coefficient consistent with the above statement is:

A. 0.8
B. 0.2
C. – 0.2
D. – 0.8

28. You are told there is a high, positive correlation between the variable ‘fitness’ and ‘hours of exercise’. The correlation coefficient consistent with the above statement is:

A. 0.3
B. 0.2
C. – 0.8
D. none of these

29. When deciding which measure of correlation to employ with a specific set of data, we should consider:

A. whether the relationship is linear or non-linear

B. the type of scale of measurement for each variable

C. both statements are true

30. The proportion of variance accounted for by the level of correlation between two variables is calculated by:

A. r

B. r^2

C. sum of x

31. To measure ranked variables the following correlation coefficient is used:

A. Pearson's

B. Spearman's

C. Fisher's

32. The estimated Pearson correlation coefficient between systolic BP (mm Hg) and age (years) in a sample of 30 middle-aged women from a given community was $r = 0.72$ ($P < 0.001$). Hence $r^2 = 0.52$. Select one statement which you believe to be true.

A. + There is substantial evidence that systolic blood pressure and age in these women are linearly related.

B. 72% of the variability of systolic blood pressure in these women can be explained by its linear relationship with age.

C. The null hypothesis that has been tested is that there is no association between systolic blood pressure and age in these women.

33. The coefficient of determination and the R-squared (R^2) are the same.

A. True

B. False

34. The estimated Pearson correlation coefficient between systolic BP (mm Hg) and age (years) in a sample of 30 middle-aged women from a given community was $r = 0.72$ ($P < 0.001$). Hence $r^2 = 0.52$. Select one statement which you believe to be true.

A. 72% of the variability of systolic blood pressure in these women can be explained by its linear relationship with age.

B. 48% of the variability of systolic blood pressure in these women is unexplained by its linear relationship with age

C. There is no true answer

35. The Pearson correlation coefficient between two variables, x and y is always positive.

A. True

B. False

36. The Pearson correlation coefficient takes the same value when the variables *x* and *y* are interchanged.

A. True

B. False

Right answers

1 – B	2 – A	3 – A
4 – A	5 – B	6 – A
7 – A	8 – B	9 – B
10 – A	11 – B	12 – B
13 – A	14 – B	15 – A
16 – A	17 – A	18 – B
19 – A	20 – B	21 – D
22 – B	23 – C	24 – C
25 – A	26 – A	27 – D
28 – D	29 – C	30 – B
31 – B	32 – A	33 – A
34 – B	35 – B	36 - A